Design & Implementation a Voice Based Camera Directional Control System

Dr.Khalifa Abboud Salim
Department of Information and Communication Engineering
Assistant prof Al-Khwarizmy College Of Engineering
Baghdad University, Baghdad, Iraq

Hayder osama
Al-Khwarizmy College Of Engineering Department of mechatronics Engineering
Baghdad University, Baghdad, Iraq

Abstract- The aim of this work is to develop an intelligent audio –video based camera tracking system. The camera rotation is controlled by active voice .Face detection and tracking was implemented using Viola-Jonse and CAMSIFT algorithms respectively. The angular position of the persons using the system was stored at the initialization phase, while the camera rotation was controlled based on the highest energy of the active user .The system was implemented using arduino microcontroller connected to pc and controlled using MATLAB and VISUAL STUDIO for video processing. And the microcontroller that use to interface the software part with hardware part.

Keywords – Audio-Video System, Viola-Jonse, CAMSHIFT.

I. Introduction

Due to the vast application using audio video in both domestic and military application in addition to the high processing capabilities of the controllers and low cost, the application of audio video signals to track objects persons are becoming essential . The integration of audio video signals has playing a very important rule for adding intelligence in the field of robotics and surveillance applications .one such application the video conference where communication between partners, customers and officers will take place. With videoconferencing becoming an essence component of Information Technology infrastructure that enables communication and joining, Companies aspiring will be considered to providers of telephony, business applications and network infrastructure services to include this ability as part of their presentation [1]. In traditional video conference systems a man is hired to recorded the whole process and thus the recorded video is done manually by control on camera movement .This type may not work properly or effectively in a room conference because there are many persons in the room that talking at the same time, secondly the cameras of conference room are mounted in a fixed point that forcing persons of conference to badly facing direction of camera vision through the meeting. In addition to, those sitting at the table and captured by camera on the screen and it is not clearly who is speaking, moreover we do not know where the person exists in the room and is it possible to change his Place. In addition to the above, the system is not secure since there is no way to check or authenticate the speaker and the man recording the conversation. Neither the system is not reliable since the camera located at the corner of the room will capture the entire members which makes the communication awkward. Mainly there are three general tracking techniques such as sensor based, vision based and microphone or voice based technique. These techniques gives the ability to camera to keep focused on the speaker only. For the sensor based [2] approaches the speaker should wear IR or magnetic sensor to send electric or magnetic signals, where the receiver unit use the signal to locate the speaker position, this means that an extra device around the neck or hand should be wear which is Annoying and uncomfortable. Vision based technique includes different approaches such as skin color [3] based tracking, motion based tracking [4, 5] or shape based [6] tracking, and when compare with other techniques like sensor based technique less annoyed but it is less accurate. Microphone arrays [7] which are adopted better to locate the audience active talking person. Microphone consider with vision is better way for sound localization and quickly approaching and unobtrusive and comfortable. This work was directed toward the design and implementation of a rotating camera based on voice energy of the speaker using array of 3 microphones.

II. Proposed Algorithm

A. Face Detection and Tracking

The Viola-Jones [8] algorithm its target is to take the frame and it scans the frame in a comprehensive manner which detect faces for any given input frame, the Viola-Jones have the ability to rescale detector for any input image which makes it close to standard approach, each time can deal with different size. Where the Viola-Jones construction of invariant detector which it takes a certain time to be fixed for calculation irrespective of image size. Viola-Jones algorithm that have the ability to building detector by using the integral image and use some simple feature called Haar feature as shown in figure (1) that have rectangular shape.

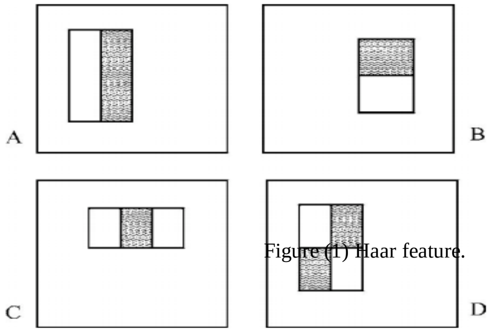


Figure (1) Haar feature.

Steerable filters, and their relatives, are excellent for detailed analysis of boundaries, image compression and texture analysis. While rectangle features are also sensitive to the presence of edges, bars, and other simple image structure, they are quite coarse. Unlike steerable filters, the only orientations available are vertical, horizontal and diagonal, Figure (2) show the face detection. The Continuously Adaptive Mean Shift Algorithm (CAMSHIFT) is an adaptation of the Mean Shift algorithm for object tracking that is intended as a step towards head and face tracking for a perceptual user interface. [9]

Figure (2) The face detection image.

It can be considered as a non-parametric technique tends to computation or climbs the gradient probability distributions in the way of implementation to find the peak of the distribution. The difference value between the center of face and center of image was calculated and used to control the motor such that the difference value is minimum.

Difference value=center of face –center of image…………… (1).

B.System Architecture

Figure (4) shows the overall system Architecture which consist of the following subsystems

a-Microphone array .

b-Microcontroller.

c-Wireless communication (Xbee1) is used for its low power short range wireless communication system .

When any person is active (talking) the corresponding signal was transmitted to the controller through (Xbee2) to initialize camera movement and start of tracking algorithm through personal computer.

The receiving system consists of the following.
1-Wireless communication (Xbee2).
2- Microcontroller interfaced to motors to control movement.
3-Pc.
4-Web cam.

At the initialization phase the camera will be rotated from 0°-180° as shown in Fig (3) such that the sitting persons are detected and save their corresponding angles. The scanning process was done for three active persons.

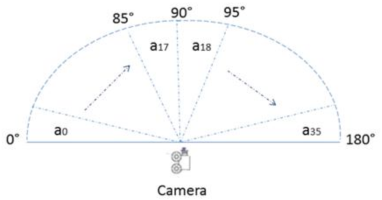
In the normal operation phase the camera movement camera can be activated based on the maximum energy received from the corresponding microphone sensor after that the process of face tracking was started until an interruption occurs with highest voice signal corresponding to the following equation.

$$E_X = \frac{1}{N}\sum_{i=0}^{N-1}|X_i|^2 \ldots\ldots\ldots\ldots\ldots (2)$$

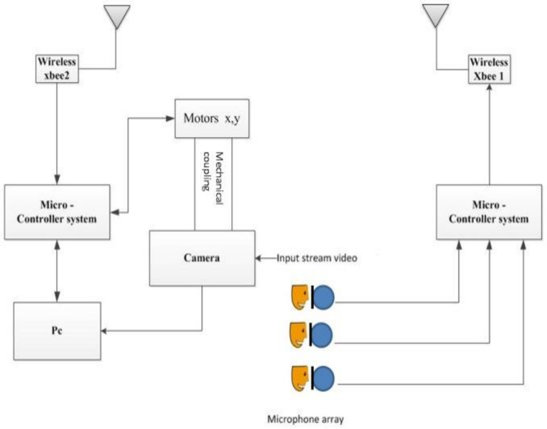
$$E_y = \frac{1}{N}\sum_{i=0}^{N-1}|y_i|^2 \ldots\ldots\ldots\ldots\ldots (3)$$

$$E_z = \frac{1}{N}\sum_{i=0}^{N-1}|z_i|^2 \ldots\ldots\ldots\ldots\ldots (4)$$

Where X_i , y_i , z_i represents the instantaneous sample time of microphone sensor (x,y,z) , and N is the number of samples taken over a time window t=4 sec which is sufficient to decide the talking person.



Figure(3). Show the camera rotation step.



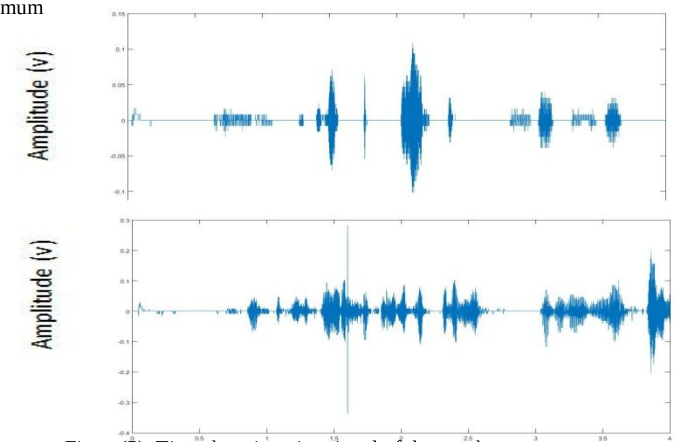
Figure(4). System Architecture.

III .PERFORMANCE EVALUATION RESULTS.

The system was developed under Matlab and Visual Studio. For any single person talk the camera will be moved toward him and when two person talk the camera will be moved toward the highest energy person. Table (1) shows the the energy sound of the first and the second person that measured through the microphone sound sensor ,while figure (5) shows the time domain voice signal of the two talking persons.

Table (1):- Show the energy level and maximum and minimum sample for first and second person.

N o S ample	S ample Range v (min)	S ample Range v (max)	I nergy $\sum\|v\|^2$
F irst person	- 0.1016	0 .1094	3 .8370
S econd person	- 0.3359	0 .2813	1 0.4005
T hird person	0	0	(



Figure(5). Time domain voice signal of the speakers.

Another test was carried where three people talk at the same time as shown figure (6).

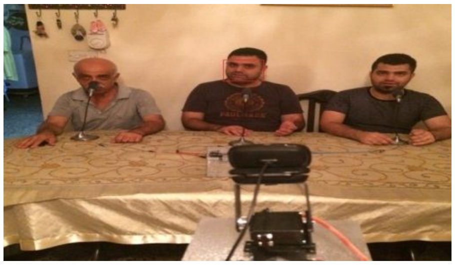
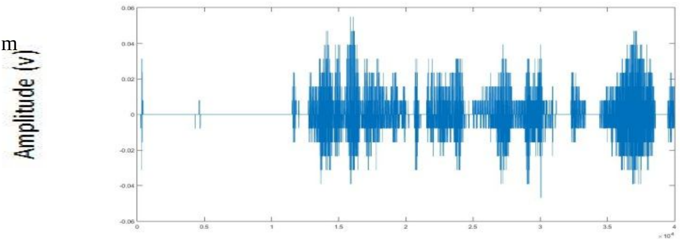


Figure (6).show the three person talk**.**

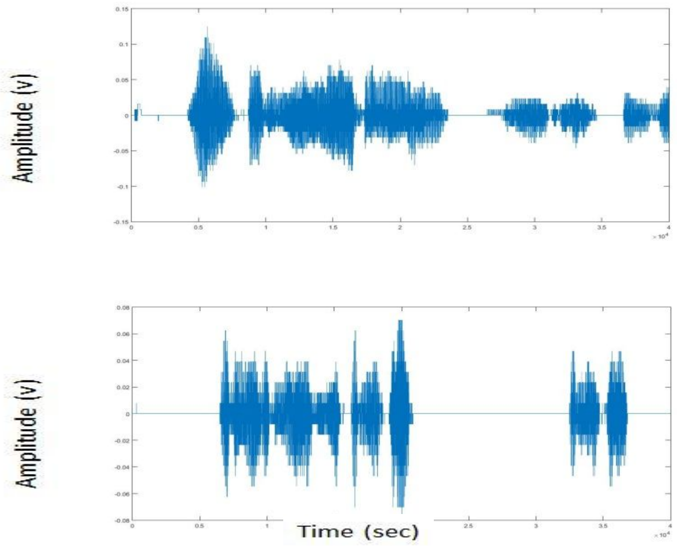
The table (2) shows the energy of the sound for three person, and figure (7) shows the time domain voice signal.

Table (2):- show the energy level and maximum and minimum Sample for three person**.**

S ample no	S ample Range v (	S ample Range v (	E nergy $\sum\|v\|^2$



	min)	**max)**	
F **irst person**	**-** **0.0469**	**0** **.0547**	**1** **.8374**
S **econd person**	**-** **0.1016**	**0** **.1250**	**8** **.3769**
T **hird person**	**-** **0.0781**	**0** **.0703**	**4** **.0550**



Figure(7). Time domain voice signal of the speakers.

IV.CONCLUSION

This work presents a design and implementation of a voice based camera tracking which uses the voice energy level to inform the system which person is speaking. The integration of voice signal and face tracking algorithm facilitate the use of such system for many automated application. The case study of video conference is proposed with such system. Many concluding remarks drawn through the investigation of the work as follows**,** Face detection is affected by many factors such as size, orientation, facial expression, illumination and occlusion. Sound localization depends on the energy from each sensor which implies the uses of voice window to determine the energy the window size in important such that the system work is stable for short period interruption**.** The servo motor gives an accurate reading of values more than the dc motor because dc motor needs gear box and servo have limit step. Sound localization integrated with camera vision insures efficient and accurate tracking system**.** Visual studio is very fast than the matlab and visual studio can be used to create many application including applications on the Android, IOS.

REFERENCE

[1] Z. Knudsen, J.Lessans, and M. Schol, Voice Tracking Camera Video Conferencing of the Future, Partial Fulfillment of the Requirement for the BSEE, May 2011.

[2] S. Mukhopadhyay, B. Smith., Passive Capture and Structuring of Lectures, Proc. of ACM Multimedia, 1999.

[3] P. Kumar G. and S. M, Real Time Detection and Tracking of Human Face Using Skin Color Segmentation and Region Properties, International Journal of Signal Processing Systems Vol. 2, No. 2, December 2014.

[4] M. K. Leung and Y. H. Yang. First sight: A human body outlines labeling system. IEEE Trans. on PAMI, 17(4):359–377, 1995.

[5] Q. Cai and J. K. Aggarwal, Tracking Human Motion Using Multiple Cameras, Computer Vision and image understanding, volume 73, issue 3, march1999.

[6] D.J. Lee, P. Zhan, A. Thomas, R. Schoenberger, "Shape-based Human Intrusion Detection", International Symposium on Defense and Security, , vol. 5438, April 12-16, 2004.

[7] Q. Liu, Y. Rui, Anoop Gupta and J. Cadiz, Automating Camera Management for Lecture Room Environments, *Special Interest Group on Computer-Human Interaction*, 2001.

[8] P. VIOLA, M.J. JONES, Robust Real-Time Face Detection, International Journal of Computer Vision, 2004.

[9] G.R.Bradski, Computer Vision Face Tracking For Use in a Perceptual User Interface, Intel Technology Journal, 1998.